Myers-Briggs Types:
A Comparison of Two Popular Measures

Stephen Bowlsby
University of British Columbia

Abstract

The Jungian-based Myers-Briggs psychological types can purportedly be measured by two independent inventories available for self-evaluation by the general public. Comparison of these tests using 23 volunteer psychology students showed reliability between tests on only two of the four dimensions. Differences between means on the two tests were significant for Introvert-Extravert ($p > .30$) and for Judging-Perceiving ($p > .10$). In addition, significant correlations between independent dimensions within tests support the lack of validity of these dependent measures, and perhaps of the theoretical construct itself. Interpretations of the results suggest directions for further study.

Myers-Briggs Types:

A Comparison of Two Popular Measures

Carl Jung's two attitudes of extraversion and introversion, and his four personality functions of thinking, feeling, sensing, intuition, judging, and perceiving, were restructured into a more simplified and systematic scheme by Isabel Myers (1962) and used as the basis of a 170-question personality inventory, the Myers-Briggs Type Indicator (MBTI), that rates an individual on eight variables that are paired to produce four dimensions: E (Extraversion) versus I (Introversion); T (Thinking) versus F (Feeling); N (Intuition) versus S (Sensing); and J (Judging) versus P (Perception). One's designation as either J or P indicates which of the other designated functions is manifested introvertedly and which extravertedly. An individual is designated as one of 16 possible personality types, each of which is given a detailed description. The test is used extensively in career counselling (The Type Reporter, 1984). I have been tested twice over 10 years, obtaining consistent and quite extreme scores, and I have found the literature based on the Myers-Briggs typography very helpful in making a critical career decision and in attempting to understand personality differences in general, particularly innate ones. Jung's original theory of the innateness of the types is maintained in all the literature based on the MBTI. Two inventories are available to the general public that purport to be testing these same variables, the Personal Style Inventory (Hogan and Champagne, 1979), and the Keirsey Temperament Sorter (Keirsey and Bates, 1984). When giving these inventories to friends, significant inconsistencies were sometimes found between tests, bringing into question the validity of these tests. It was also noted recently that contemporary research on personality type does not refer to Jung's functions at all, only to the generally accepted attitudes of extroversion/introversion (McCrae and Costa, 1987), and so the

validity of the Myers-Briggs construct itself was also called into question. A comparison of the reliability of scores between the two publicly-available tests could indicate these tests' validity as dependent measures of the MBTI variables, and could perhaps even shed light on the construct validity of the Myers-Brigg's and Jungian theory itself, or at least for measures of it that depend on self-report. In short, this study was done to evaluate the degree of caution advisable in relying on the results of either of these tests.

In comparing individuals' responses on the two tests, I predicted that the I-E dimension (extraversion) would be most reliable, along with the N-S dimension because of the similarity of Jungian Intuition (N) with the Openness factor of the generally-accepted five-factor model described by McCrae and Costa (1987). The predicted reliability of I-E was implied by the Keirsey test itself, in that it included only half as many questions for this dimension. Based on comments from people taking these tests, I also expected to find the lowest reliability on the T-F dimension.

Method

Subjects

Subjects were selected from two intersessional night classes in psychology at The University of British Columbia. One class was for a required course in experimental methodology, and had a high proportion of mature students. As a member of this class myself, I encountered these students in every class session, and so to encourage volunteers I gave them the option of including their name on the test so that I could interpret their results for them if they so desired. The entire class of 19 received the test, as did the class Teaching Assistant, a graduate student in personality psychology who had previously taken the Myers-Briggs test. Fifteen tests were returned, 4 with requests for interpretations. In the second class, an introductory course in

physiological psychology, I emphasized the personal-interest factor as a motivator to volunteer, while allowing them all to remain anonymous by including a self-scoring sheet and a phone number they could call to have the scores interpreted. Twenty students volunteered and received tests; nine were returned, of which one was discarded because it was filled out incorrectly. None of these students called for interpretations. Total subjects were 23.

Measures

Test 1 is the Personal Style Inventory, which includes 32 questions, eight for each dimension, each in the form of a graded scale of preference between two phrases, where the sum of integer values given to each member of the pair must equal five. The Myers-Briggs construct is thus incorporated into the pair choice, in that scores are automatically inversely proportional; total scores on a given variable can be directly deduced from the total scores on the related variable (e.g., introversion score deduced from extraversion score). This logic applies also to Test 2, the Keirsey Temperament Sorter, which includes 70 forced-choice questions.

Design and Procedure

Student volunteers were allowed a week to complete and return the two tests. Sequencing effects were counterbalanced by stapling the two tests together, half of them in reversed order, and requesting that subjects fill out the tests in the order that they appeared. Gender was not controlled for since statistics show that mean preferences for each sex are the same, except for a negligible difference on the Thinking-Feeling dimension (Keirsey, 1984).

Four scores were computed for each individual on both tests by taking the total score on one of the variable-pairs only for each dimension, yielding scores for Extraversion (E1 for test 1, E2

for test 2), Intuition (N1, N2), Thinking (T1, T2), and Judging (J1, J2). These four variables thus indicate preferences on each of the four dimensions. In order to compare means for each dimension between tests, the raw scores on test 2 were linearly transformed (N2, T2, and J2 multiplied by four, and E2 by two, to produce a scale from zero to 40 for all variables). Between tests, correlations were computed and the means compared for the four dimensions (note opposite null hypotheses: that correlation is zero; that the means are the same). Within tests, correlations between dimensions, which should not be correlated according to the Myers-Briggs theory, were also investigated.

Results

Between-Tests Reliability

Table 1 shows the results of the between-tests comparisons of the four dimensions. All dimensions are significantly correlated across tests ($p < .01$), but only two dimensions are reliable. The Thinking-Feeling dimension is the most reliable and the most highly correlated ($r = .82$; $t = 0.97$, $p > .30$), followed by the Sensing-Intuition dimension ($r = .73$; $t = -1.64$, $p > .10$). On the other hand, Extraversion-Introversion ($t = -2.73$, $p < .05$) and Judging-Perceiving ($t = -3.28$, $p < .01$) do not appear to be measuring the same thing on both tests. The restricted range of J1, as shown in Table 2, which was not found to be due to any computational error, would reduce r and p for Judging-Perceiving; the reliability of this dimension may therefore be closer to that of Extraversion-Introversion.

Insert Table 1 about here

Insert Table 2 about here

Other statistical factors were examined to detect any
confounding effects on this between-tests comparison. As shown
in Table 2, means for all dimensions on both tests are roughly
the same, and are at about the midpoint of the 40-point scale,
indicating roughly equal mean proportions of all eight variables.
A mild difference in the Thinking score indicates a slight
population preference for Feeling. Skewness (.16 to -.64) is not
great enough to significantly affect the results. Standard
deviations on test 2 are consistently higher than on test 1, but
using a one-sample chi-square test comparing the mean variance on
each test to the between-test mean variance[1], the difference was
not found to be significant (X^2_{22} = 33.7, p > .10).
Additionally, the difference in variances is likely to be due to
scoring methods rather than differential test sensitivities.

Reliable dimensions are T-F and S-N. Unreliable are E-I and
J-P.

Within-Test Correlations

Table 3 shows correlations between dimensions within each
test. The two variables shown in each pair of correlations come
from two different dimensions for which there is not a "forced"
correlation due to the inherent constructions of the tests, as is
the case for the two paired variables within each dimension. The
Myers-Briggs theory would predict no significant correlation for
any of these variables from separate dimension pairs. When
considering the following data, recall that the Myers-Briggs
dimensions are E-I, N-S, T-F, and J-P -- only one of each
variable-pair is used in this table to represent the dimension
(E, N, T, and J). For clarity, when a negative correlation is
found, the opposite variable of the dimension-pair will be
referred to in the description and later discussion so that all
correlations referred to are positive (e.g., a negative
correlation between Extraversion [E] and Thinking [T] will be
referred to as a positive correlation between E and F [Feeling]).

Insert Table 3 about here

Only one pair -- E and N -- is clearly free from correlation on both tests (test 1: r = .006, p > .95; test 2: r = .23, p > .25). Two other pairs are clearly correlated (i.e., positively) on both tests: T-J, and N-F[2] (r > .40, p < .02). The remaining three pairs appear to show different correlations in each test. To test the significance of each of these differences, one-sample z-tests were performed using Fisher-Z transformations and a null hypothesis taking one r as the parameter and the other r as the sample statistic[3]. Results showed that correlation differences between tests on E-T and E-J pairs are not significant (z < |0.95|, p > .30). The remaining pair, N-J, is the only one showing a significant difference between tests (z > 2.0, p < .01).

In summary:

Uncorrelated: E-I.

Positively correlated: I-F, T-J, E-P, and E-F.

Test difference: N-P correlated on Test 2 only (r =.70).

Discussion

At least one of these two inventories is probably not a valid measure of the MBTI construct, and results fail to confirm the validity of the construct itself. Against prediction, the extraversion measure showed the lowest reliability and the thinking-feeling measure showed the highest, while, consistent with prediction, the intuition-sensing measure was reliable. Correlations between dimensions within each test imply very poor validity of dependent measures, or indeed very poor validity of the theory itself. Discrepancies can be due to any combination of three main factors: invalid construct, invalid measures, or unreliable self-evaluation.

Measures and Construct

Attempted interpretation of the extraversion discrepancy revealed the fact that modern definitions of this variable are not Jung's. Jung's distinction was that extraverts were "conditioned more by the objects of their interest", while introverts were " conditioned more by their own inner self" (Jung, 1959, Ed: Laszlo). In contrast, accepted modern definitions stress lively sociability, activity level, assertiveness, sensation-seeking, and particularly surgency (dominance and activity). Examining the test questions shows that they are in fact based on this accepted definition, so here the lack of reliability is not likely due to the construct but to the measure. In fact, Jung's objective-subjective distinction is shown more in the T-F scale, and perhaps adds to the validity of that dimension. Possible confounds discovered by close investigation of test questions support most of the inter-dimension correlations.

Extraversion: Three out of eight E-I questions on Test 1 are judgment-oriented, while none on Test 2 are, and Test 2 (Keirsey) E-I questions seem to employ more of the modern range of extravert discriptors, so it may be the more valid measure. On both, however, the sociability aspects relate to Feeling. The correlation of E with P but not with N implies an E-S correlation, but inspection of the questions failed to support this.

Judging-Perceiving: On Test 1, three of eight questions appear to imply surgency[4], and three more imply thinking[5], while on Test 2 the ratio is 3/20[6] and 5/20[7]; so both are confounded, but Test 2 (Keirsey) is less so. The N-P correlation found only for Test 2 is evident in 5/10 of the J-P questions, which stress N-type functions as opposed to S-type[8]. Finally, the correlation found between N and F, which are perceiving and judging functions respectively, could confound this dimension.

Self-Report

Some of the unreliability may be due to subjective clouding of judgment, where "in every pronounced type there exists a special tendency towards compensation for the onesidedness of the type" (Jung, 1959), and where, for example, a function seen introvertedly will seem very different from the same function extraverted. Noting, for example, that the unreliable E correlates with reliable S and F, perhaps "SF" people are less able to identify their E-I position.

Positive self-presentation may play a role also, even if unconsciously, particularly in response to social discrimination against introversion and "judging". However, such discrimination would be expected against thinking as well, so perhaps the socialization aspect of Feeling is not reflected in the questions as much as the subjective-objective dichotomy, which has more equal numbers of proponents in society. Thus J-P and E-I could be obscured by conformity, while in this test T-F and N-S are not. An alternative interpretation of the T-F reliability comes from the observation that T-F questions do seem to include a strong "compassion" component -- the result might be a strong self-representation bias for F, which would produce the observed reduction in mean T score and the illusion of higher reliability.

Cases and Opinion

The Teaching Assistant was an "ENTJ" on the MBTI, an "ENFJ" on Test 2, and "ESFJ" on Test 1. His scores were not high on any given variable, and I am certain from personal observation and knowledge of the type descriptions that the ESFJ designation is wrong. In another case, a student was consistently very high on "INT" and equal on the J-P dimension across tests. My impression is that these tests may be meaningful only for people who score very high or completely neutral on a given dimension, and I suspect that these people would just as reliably, and more

easily, be able to select their type (or lack thereof) from overt descriptions of the main variables.

Because of the built-in construct of the question pairs, a revealing test of construct validity would be to separate the paired questions, randomly mix them, and have them ranked independently to see how correlations compare to the results of this study.

References

Hogan, R.C., Craig, W.C. (1979). _Supervisory and Management Skills: A Competency Based Training Program for Middle Managers of Educational Systems_. Published privately.

Jung, C.G. (1959). _The Basic Writings of C.G. Jung._ Edited by V. De Laszlo. New York: Random House.

Keirsey, D., Bates, M. (1984). _Please Understand Me: Character and Temperament Types_. Del Mar California: Gnosology Books Ltd.

McCrae, R. R., Costa Jr., P.T. (1987). Validation of the Five-Factor Model of Personality Across Instruments and Observers. _Journal of Personality and Social Psychology,_ _1_, 81-90

Myers, I. (1962). _Manual: The Myers-Briggs Type Indicator_. Palo Alto, California: Consulting Psychologists Press.

The Type Reporter: A Quarterly Newsletter on Psychological Type (1984). Alexandria, Virginia.

Author Notes

Footnotes

[1]A direct test on correlated variances could not be done because it is not taught in Psych 316.

[2]From negative correlation of N-T pairs.

[3]Direct testing of dependent correlation coefficients could not be done because it is not taught in Psych 316.

[4]4, 12, 28

[5]8, 16, 20

[6]55, 21, 27

[7]56, 70, 42, 34, 7

[8]69, 48, 34, 28, 63

Table 2

Statistics for Each Variable

NUMBER OF VALID OBSERVATIONS (LISTWISE) =   23.00

VARIABLE	MEAN	S.E. MEAN	STD DEV	SKEWNESS	S.E. SKEW	RANGE	LABEL
E1	19.696	1.104	5.295	-.638	.481	20.000	Extroversion on Mgmt Inv
N1	21.478	1.285	6.163	.294	.481	23.000	Intuition on Mgmt Inv
T1	18.522	1.494	7.166	.231	.481	28.000	Thinking on Mgmt Inv
J1	21.478	.998	4.785	-.444	.481	16.000	Judging on Mgmt Inv
E2	24.348	2.162	10.369	-.413	.481	36.000	Extroversion on Kiersey
N2	23.913	2.128	10.207	-.260	.481	36.000	Intuition on Kiersey
T2	17.478	1.883	9.030	.158	.481	32.000	Thinking on Kiersey
J2	25.826	1.639	7.860	.233	.481	26.000	Judging on Kiersey

Table 1

Comparison of Dimensions Between Tests

- T-TEST - - - - - - - - - - - - - - - - - -

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR	*(DIFFERENCE) MEAN	STANDARD DEVIATION	* CORR.	2-TAIL PROB.	T VALUE	2-TAIL PROB.
E1	Extroversion on Mgmt Inv									
		19.6957	5.295	1.104						
	23				-4.6522	8.183	0.625	0.001	-2.73	0.012
		24.3478	10.369	2.162						
E2	Extroversion on Kiersey									
N1	Intuition on Mgmt Inv									
		21.4783	6.163	1.285						
	23				-2.4348	7.115	0.728	0.000	-1.64	0.115
		23.9130	10.207	2.128						
N2	Intuition on Kiersey									
T1	Thinking on Mgmt Inv									
		18.5217	7.166	1.494						
	23				1.0435	5.183	0.819	0.000	0.97	0.345
		17.4783	9.030	1.883						
T2	Thinking on Kiersey									
J1	Judging on Mgmt Inv									
		21.4783	4.785	0.998						
	23				-4.3478	6.365	0.587	0.003	-3.28	0.003
		25.8261	7.860	1.639						
J2	Judging on Kiersey									

Table 3

Correlations Between Dimensions Within Tests

- - - - - - - - - - - - - - - P E A R S O N   C O R R E L A T I O N   C O E F F I C I E N T S - -

VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR	
E1 WITH N1	.0061 N(23) SIG .978	E1 WITH T1	-.4580 N(23) SIG .028	E1 WITH J1	-.2075 N(23) SIG .342	N1 WITH T1	-.5081 N(23) SIG .013	N1 WITH J1	-.2732 N(23) SIG .207
E2 WITH N2	.2305 N(23) SIG .290	E2 WITH T2	-.2699 N(23) SIG .213	E2 WITH J2	-.4142 N(23) SIG .049	N2 WITH T2	-.4306 N(23) SIG .040	N2 WITH J2	-.7027 N(23) SIG .000

- - - - - - - - - -

VARIABLE PAIR	
T1 WITH J1	.5478 N(23) SIG .007
T2 WITH J2	.4982 N(23) SIG .016